INDIAN INSTITUTE OF SPACE SCIENCE AND TECHNOLOGY
THIRUVANANTHAPURAM

Assignment #1

Due on 28-08-2014

SUHAS S
(SC14M081)

Contents

1.History of Data Mining

The Data Mining field proposes the development of methods and techniques for assigning useful meanings for data stored in databases. It gathers researches from many study fields like machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, data visualization and grids.**Data Mining represents a set of specific algorithms of finding useful meanings in stored data**[1]. The DM process has developed due to the immense volume of data that must be handled easier in areas such as: business, medical industry, astronomy, genetics or banking field.

Data mining has become an established discipline within the scope of computer science.Broadly data mining can be described as a set of mechanisms and techniques, realized in software, to extract **hidden information from data**[2]. By the early 1990s, data mining was commonly recognized as a subprocess within a larger process called knowledge discovery in databases or KDD.

The data that data mining techniques were **originally directed at was tabular data**[1] and, given the processing power available at the time,computational efficiency (and particular the number databases accesses) was of significant concern. As the amount of processing power generally available increased, processing time (although still an issue) became less of a concern and was replaced with a desire for accuracy and a desire to mine ever larger data collections.

When the DM term was introduced back in 1989 by the researcher Gregory Piatesky Shapiro, there were not too many data mining instruments for resolving one single task. A good example is the C4.5 decision tree algorithm (Quinlan, 1986) and SNNS neural network, or parallel-coordinate visualization (Inselberg, 1985). These tools were hard to use and required important data preparation (Piatesky-Saphiro, 1991)[1].

The popularity of data mining increased significantly in the 1990s, notably with the establishment of a number of dedicated conferences.It became common place for commercial enterprises to maintain data in computer readable form, in most cases this was primarily to support commercial activities, **the idea that this data could be mined often came second**[2]. The 1990s also saw the introduction of customer loyalty cards (particularly with respect to large super market chains) that allowed enterprises to record customer purchases, **the resulting data could then be mined to identify customer purchasing patterns**[2]. The popularity of data mining has continued to grow over the last decade with a particular current emphasis on mining non-standard data (i.e. non-tabular data).

The second-generation data mining systems were called suites[1] and were developed by vendors, starting from 1995. These tools took into account that the DM process requires multiple types of data analysis, and most of the effort is spent in the data cleaning and preprocessing steps. Suites like SPSS Clementine, SGI Mineset, IBM Intelligent Miner, or SAS Enterprise Miner allowed the user to perform several discovery tasks (usually classification, clustering, and visualization) and also supported data transformation and visualization. One of the most important advances, pioneered by Clementine, was a GUI (Graphical User Interface) that allowed users to build their knowledge discovery process visually (Piatesky-Saphiro, 1991).

By the year 1999, there were over 200 tools available for solving different tasks but even the best of them addressed only a part from the overall DM framework. Data still had to be cleaned and preprocessed. The development of this type of applications in areas like direct marketing, telecom, and fraud detection, led to emergence of data mining based ”vertical solutions”. The best examples of such applications are the **systems HNC Falcon for credit card fraud detection**[1], IBM Advanced Scout for sports analysis and NASD DM Detection system (Kirkland, 1999;Piatesky-Saphiro, 1991).

A very important issue is the way that data was stored over the time. Many years the main approach was to use a

specific DM method or algorithm on a data set. In most cases the data set was stored in a centralized database. In present, because of big volumes of data the main solution is to use distributed databases systems. For mining in this data in the traditional way it is supposed that all data stored on local computers should be transferred on a central point for processing. In most cases this would be impossible because the existent connection bandwidth won't permit such big transfers. A very important matter is that when big transfers are made over the Internet can appear security issues: the intimacy of client's data must be kept. Thus, a new DM study area appeared that was called **Privacy Preserving Data Mining**[1].This field focuses on studying the security risks that can occur in the DM process. Because the number of steps that are included in the DM framework is relative big client's data that are mined can be violated. Privacy Preserving Data Mining tries to create algorithms that may prevent such problems.

In the last years the DM process was approached from two perspectives: parallel and distributed computing. These directions led to the apparition of Parallel DM and Distributed DM[1]. In Parallel DM, data sets are assigned to high performance multi-computer machines for analysis. The availability of this kind of machines is increasing and all algorithms that were used on single-processor units must be scaled in order to run on parallel-computers. The Parallel DM technology is suitable for scientific simulation, transaction data or telecom data. Distributed DM must provide solutions for local analysis of data and global solutions for recombining local results from each computing unit without causing massive data transfer to a central server. Parallel computing and distributed DM are both integrated in Grid technologies.

The evolution of Data Mining can be summarized best with the following table[3]

Evolutionary Step	**Business Question**
Data Collection (1960s)	What was my total revenue in the last five years?
Data Access (1980s)	What were unit sales in New England last March?
Data Warehousing and Decision Support	What were unit sales in New England last March? Drill down to Boston.
Data Mining	What's likely to happen to Boston unit sales next month? Why?

2.Real World Applications of Data Mining

Application domain of Data Mining consists[4]

- Analytics
- Bioinformatics
- Business Intelligence
- Data analysis
- Data warehouse
- Decision support system
- Drug discovery
- Predictive Analysis
- Web Mining etc

Some of the real world applications of Data Mining are

2.1 Data Mining for Financial Data Analysis

In the banking industry, data mining is used heavily in the areas of **modeling and predicting credit fraud, in evaluating risk, in performing trend analyses, in analyzing profi tability, as well as in helping with direct-marketing campaigns**. In the financial markets, neural networks have been used in forecasting stock prices, options trading, rating bonds, portfolio management, commodity-price prediction, and mergers and acquisitions analyses; it has also been used in forecasting financial disasters. Daiwa Securities, NEC Corporation, Carl & Associates, LBS Capital Management, Walkrich Investment Advisors, and O'Sallivan Brothers Investments are only a few of the financial companies who use neural network technology for data mining. A wide range of successful business applications has been reported, although the retrieval of technical details is not always easy[5].

The widespread use of data mining in banking has not been unnoticed. Bank Systems & Technology commented that data mining was the most important application in financial services in 1996. For example, fraud costs industries billions of dollars, so it is not surprising to see that systems have been developed to combat fraudulent activities in such areas as credit card, stock market, and other fi nancial transactions. Fraud is an extremely serious problem for credit card companies. For example, Visa and MasterCard lost over $700 million in 1995 from fraud. A neural network-based credit card fraud-detection system implemented in Capital One has been able to cut the company's losses from fraud by more than 50%.

2.2 Data Mining for the Telecommunications Industry

The hypercompetitive nature of this industry has created a need to understand customers, to keep them, and to model effective ways to market new products. This creates a great demand for data mining to help understand the new business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of services. In general, the telecommunications industry is interested in answering some strategic questions through data mining applications such as:

- How does one retain customers and keep them loyal as competitors offer special offers and reduced rates?
- Which customers are most likely to churn?
- What characteristics indicate high-risk investments, such as investing in new fiber-optic lines?
- How does one predict whether customers will buy additional products like cellular services, call waiting, or basic services?
- What characteristics differentiate our products from those of our competitors?

Companies like AT&T, AirTouch Communications, and AMS Mobile Communication Industry Group have announced the use of data mining to improve their marketing activities. There are several companies including Lightbridge and Verizon that use data mining technology to look at cellular fraud for the telecommunications industry. Another trend has been to use advanced visualization techniques to model and analyze wireless telecommunication networks[5].

2.3 Data Mining for the Retail Industry

Slim margins have pushed retailers into data warehousing earlier than other industries. Retailers have seen improved decision - support processes leading directly to improved efficiency in inventory management and financial forecasting. The early adoption of data warehousing by retailers has allowed them a better opportunity to take advantage of data mining. The retail industry is a major application area for data mining since it collects huge amounts of data on sales, customer-shopping history, goods transportation, consumption patterns, and service records, and so on. The quantity of data collected continues to expand rapidly, especially due to the increasing availability and popularity of business conducted on the Web, or e-commerce. Today, many stores also have Web

sites where customers can make purchases online. A variety of sources and types of retail data provide a rich source for data mining[5].

Retail data mining can help identify customer-buying behaviors, discover customer shopping patterns and trends, improve the quality of customer services, achieve better customer retention and satisfaction, enhance goods consumption, design more effective goods transportation and distribution policies, and, in general, reduce the cost of business and increase profitability. In the forefront of applications that have been adopted by the retail industry are direct-marketing applications. The direct-mailing industry is an area where data mining is widely used. Almost every type of retailer uses direct marketing, including catalogers, consumer retail chains, grocers, publishers, B2B marketers, and packaged goods manufacturers. **The claim could be made that every Fortune 500 company has used some level of data mining in their direct-marketing campaigns**[5].

Large retail chains and groceries stores use vast amounts of sale data that are "information-rich." Direct marketers are mainly concerned about customer segmentation, which is a clustering or classification problem.

Retailers are interested in creating data mining models to answer questions such as:

- What are the best types of advertisements to reach certain segments of customers?
- What is the optimal timing at which to send mailers?
- What is the latest product trend?
- What types of products can be sold together?
- How does one retain profitable customers?
- What are the significant customer segments that buy products?

Data mining helps to model and identify the traits of profitable customers, and it also helps to reveal the "**hidden relationship**" in data that standard-query processes have not found. IBM has used data mining for several retailers to analyze shopping patterns within stores based on point-of-sale (POS) information. For example, one retail company with $2 billion in revenue, 300,000 UPC codes, and 129 stores in 15 states found some interesting results: " . . we found that people who were coming into the shop gravitated to the left hand side of the store for promotional items, and they were not necessarily shopping the whole store." Such information is used to change promotional activities and provide a better understanding of how to lay out a store in order to optimize sales.

2.4 Data Mining in Health Care and Biomedical Research

With the amount of information and issues in the health-care industry, not to mention the pharmaceutical industry and biomedical research, opportunities for data mining applications are extremely widespread, and benefits from the results are enormous. Storing patients records in electronic format and the development in medical information systems cause a large amount of clinical data to be available online. Regularities, trends, and surprising events extracted from these data by data mining methods are important in assisting clinicians to make informed decisions, thereby improving health services.

Data mining has been used extensively in the medical industry. Data visualization and artificial neural networks are especially important areas of data mining applicable in the medical field. For example, NeuroMedicalSystems used neural networks to perform a pap smear diagnostic aid. Vysis Company uses neural networks to perform protein analyses for drug development. The University of Rochester Cancer Center and the Oxford Transplant Center use KnowledgeSeeker, a decision tree based technology, to help with their research in oncology.

The past decade has seen an explosive growth in biomedical research, ranging from the development of new pharmaceuticals and advances in cancer therapies to the identification and study of the human genome. The logic behind investigating the genetic causes of diseases is that once the molecular bases of diseases are known, precisely targeted medical interventions for diagnostics, prevention, and treatment of the disease themselves can be developed. Much of the work occurs in the context of the development of new pharmaceutical products that can be used to fight a host of diseases ranging from various cancers to degenerative disorders such as Alzheimer's Disease.

A great deal of biomedical research has focused on DNA-data analysis, and the results have led to the discovery of genetic causes for many diseases and disabilities. An important focus in genome research is the study of DNA sequences since such sequences form the foundation of the genetic codes of all living organisms. If we understand DNA sequences, theoretically, we will be able to identify and predict faults, weaknesses, or other factors in our genes that can affect our lives. **Getting a better grasp of DNA sequences could potentially lead to improved procedures to treat cancer, birth defects, and other pathological processes. Data mining technologies are only one weapon in the arsenal used to understand these types of data**, and the use of visualization and classification techniques is playing a crucial role in these activities[5].

2.5 Data Mining in Science and Engineering

Enormous amounts of data have been generated in science and engineering, for example, in cosmology, molecular biology, and chemical engineering. In cosmology, advanced computational tools are needed to help astronomers understand the origin of large-scale cosmological structures as well as the formation and evolution of their astrophysical components (galaxies, quasars, and clusters). Over 3 terabytes of image data have been collected by the Digital Palomar Observatory Sky Survey, which contain on the order of 2 billion sky objects. It has been a challenging task for astronomers to catalog the entire data set, that is, a record of the sky location of each object and its corresponding classification such as a star or a galaxy. The Sky Image Cataloguing and Analysis Tool (SKICAT) has been developed to automate this task. The SKICAT system integrates methods from machine learning, image processing, classification, and databases, and it is reported to be able to classify objects, replacing visual classification, with high accuracy[5].

In molecular biology, recent technological advances are applied in such areas as molecular genetics, protein sequencing, and macro-molecular structure determination. Artificial neural networks and some advanced statistical methods have shown particular promise in these applications. In chemical engineering, advanced models have been used to describe the interaction among various chemical processes, and also new tools have been developed to obtain a visualization of these structures and processes. Pavilion Technologies Process Insights, an application development tool that combines neural networks, fuzzy logic, and statistical methods has been successfully used by Eastman Kodak and other companies to develop chemical manufacturing and control applications to reduce waste, improve product quality, and increase plant throughput. Historical process data is used to build a predictive model of plant behavior and this model is then used to change the control set points in the plant for optimization[5].

DataEnginee is another data mining tool that has been used in a wide range of engineering applications, especially in the process industry. The basic components of the tool are neural networks, fuzzy logic, and advanced graphical user interfaces. The tool has been applied to process analysis in the chemical, steel, and rubber industries, resulting in a saving in input materials and improvements in quality and productivity.

This is certainly not an inclusive list of all data mining activities, but it does provide examples of how data mining technology is employed today. We expect that new generations of data mining tools and methodologies will increase and extend the spectrum of application domains.

References

[1] Jing He, *Advances in Data Mining: History and Future*. Economics and Management School, Wuhan University.

[2] Frans Coenen, *Data mining: past, present and future*. Department of Computer Science, The University of Liverpool.

[3] http://www.laits.utexas.edu/ anorman/BUS.FOR/course.mat/Alex/

[4] Data Mining wikipedia - http://en.wikipedia.org/wiki/Data_mining

[5] Mehmed Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition*. Appendix B Data Mining Applications